MECHANISMS BEFORE METHODS: MEASURING HOW AI DESIGNS AI RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Can an AI system do more than propose a plausible research idea—can it choose a principle worth pursuing and turn it into a workable method? Existing evaluations either score open-ended proposals, where reliable negative examples are hard to obtain, or end-to-end execution, where scientific judgment is mixed with engineering. We introduce *Mechanisms Before Methods* (MBM), a benchmark that scores research design before any experiment is run, built from 789 papers. MBM represents each method by its *mechanism*: the operating principle that explains how it works without revealing its paper-specific implementation. Each paper supports two paired tasks. *Mechanism Allocation* chooses the documented route among plausible alternatives; *Mechanism Realization* works out how that route should operate on the target problem. Across ten systems the two abilities rise together ($r = 0.857$), while within a single paper a correct choice adds under one point of realization quality. Asking a model to draft its own realization before choosing lowers its accuracy, while prewritten realizations raise it by 4.70 points in a matched diagnostic: a plausible plan is not evidence for a choice. Reinforcement learning improves Realization F1 by 3.37 points, and the ability transfers: a short Realization phase improves later Allocation itself. MBM makes pre-experimental research decisions separately measurable, and suggests that learning how a principle works helps decide whether it is worth pursuing.

1 INTRODUCTION

The most consequential decisions in research often precede experimentation. Given a problem, a researcher must decide which operating principle deserves investment, then work out how that principle should act on the problem's objects and constraints. The first chooses where to invest; the second turns that choice into a method. AI systems increasingly write the code and run the experiments for a research project, yet we still lack a controlled way to tell whether they can make the design decisions that come first.

Existing evaluations face a construction problem. Free-form ideation appears direct, but supplies no dependable source of negative examples: a route not chosen by the original authors may still be viable, while generated distractors can introduce stylistic shortcuts. End-to-end reproduction moves to the other extreme, mixing route choice with implementation, engineering, and experimental outcomes. A single proposal score inherits both problems: it cannot distinguish choosing the wrong principle from failing to make the right principle concrete.

Published papers provide authentic problems and methods, but presenting the full method gives away the answer. We therefore abstract each method to its *mechanism*: the operating principle that explains how the method produces its effect, stripped of paper-specific bindings and pipeline details. A mechanism says more than a topic label because it constrains the form of a solution, but less than a complete method, leaving the model to work out how the principle applies.

Complete methods are also poor units for comparison. A paper usually bundles an architecture with an objective and a training recipe, so comparing two methods leaves unclear which scientific choice is being tested. Mechanisms put candidates in a common language: each makes a distinct claim about what should produce the desired effect. The source paper supplies an auditable positive—the route its authors actually pursued—and the other cards are mechanisms that other papers pursued, not inventions of a generator.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Mechanisms Before Methods (MBM) uses this intermediate representation to create two paired views of every paper:

$$\textsc{Allocation} : x \to m, \qquad \textsc{Realization} : (x, m) \to r. \tag{1}$$

Mechanism Allocation asks which mechanism matches the source-documented research route. Mechanism Realization supplies that mechanism and asks how it should operate on the target problem. Realization is evaluated with the documented mechanism, so an Allocation error does not carry into the second task. Both are scored before any experiment is run, which is what separates MBM from end-to-end reproduction. We can then supply, perturb, or train realizations and ask whether they change the model's choice.

Prewritten realizations help models choose mechanisms, but their own drafts do not: a plausible plan is not evidence for a choice. Stronger systems do well on both tasks, yet the two abilities break apart on the same paper. The link is at least partly learnable, and it runs in one direction: training Realization improves Allocation itself, which makes realization a candidate prerequisite skill and not a second scoreboard.

The resulting benchmark pairs 789 papers against a reviewed library of 243 mechanism cards. It measures route choice and realization separately, while keeping them linked closely enough to test when one helps the other.

2 Research design as linked decisions

Let x_t describe a problem, including its objective, available information, and constraints; m_t an operational mechanism; r_t its target-specific realization; and e_t evidence from execution, derivation, or observation. Their forward dependencies are

$$x_t \longrightarrow m_t, \qquad (x_t, m_t) \longrightarrow r_t, \qquad (x_t, m_t, r_t) \longrightarrow e_t. \tag{2}$$

Evidence may revise the problem, mechanism, or realization, producing the familiar research loop. The loop need not follow a rigid chronology. MBM measures the middle two transitions while holding execution outside the scoring target.

The abstraction must hide paper identity and target-specific implementation while preserving enough operational structure to constrain a method and support source-grounded evaluation. Figure 1 shows what each candidate level does to the space of admissible methods. Mechanisms are also the level at which research routes become comparable. Two methods may differ in architecture and training recipe while making the same underlying scientific bet; conversely, similar components may serve different principles. What transfers is the relational structure of the bet, not the attributes of the objects it concerns (Gentner, 1983). Scoring a distractor as incorrect does not assert that the mechanism would have failed; it records that the model did not identify the route this paper took.

A paper supplies a documented instance $d_i = (x_i, m_i^*, r_i^*, e_i^*)$, where the asterisk marks the route recorded in the source. Given four reviewed candidate cards $\mathcal{C}_i$ containing m_i^*, MBM constructs

$$\textsc{Allocation} : (x_i, \mathcal{C}_i) \longrightarrow \text{ranking of } \mathcal{C}_i, \qquad \text{reference } m_i^*, \tag{3}$$

$$\textsc{Realization} : (x_i, m_i^*) \longrightarrow \widehat{r}_i, \qquad \text{reference requirements from } r_i^*. \tag{4}$$

Mechanism Realization receives the documented mechanism, not the allocation output. The two views therefore share a research context without propagating an upstream error.

The reference earns its standing, though not by being optimal: it is the route a research group committed to under real constraints and then reported as working, in enough detail to audit afterwards. Generated distractors have no comparable standing—no one ever chose them—and retrospective ratings give an opinion about a route, not a decision to pursue it. Other routes may well have been viable.

Allocation and Realization sit on opposite sides of the same research commitment. Allocation asks where to invest among plausible principles; Realization asks whether a supplied principle is understood well enough to specify what must happen in the target setting. A model may select a familiar card without understanding its consequences, or competently realize a supplied card without recognizing when it should be chosen. We separate the two views in order to measure them, not to claim that a researcher performs them in sequence: choosing a route may well involve working out its consequences, and MBM makes that interaction testable by varying the realization in one view while fixing the mechanism in the other (Jacobs & Wallach, 2021; Bowman & Dahl, 2021).

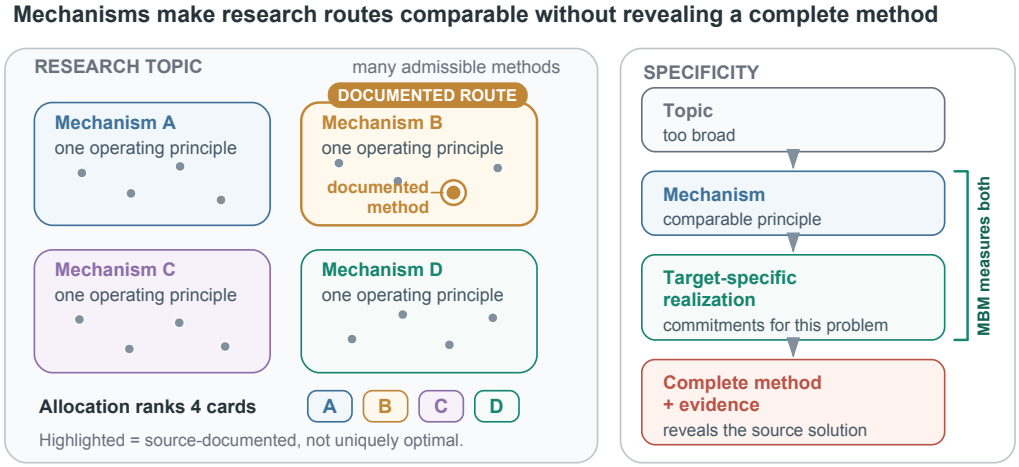

Figure 1: What each abstraction level does to the space of admissible methods. A topic admits too many methods to compare; a complete method names a single point. A mechanism admits many methods, so naming one does not give the source solution away, while four mechanisms under the same topic stay comparable—which is exactly an Allocation item. MBM measures the middle two levels of the ladder. Positions are illustrative, and the wording paraphrases the public description of LanczosNet (Liao et al., 2019).

3 Related Work

Scoring research by what it produces. The AI Scientist automates a path from ideation through experiments and paper writing, and its successor produced a workshop paper accepted under peer review with no human-authored code template (Lu et al., 2026; Yamada et al., 2025). Benchmarks in this line score what an agent achieves after code runs: MLAgentBench and ScienceAgentBench on executable machine-learning and data-driven discovery tasks (Huang et al., 2024b; Chen et al., 2025), DiscoveryBench on multi-step analysis workflows over provided datasets (Majumder et al., 2024), MLE-bench on 75 Kaggle competitions against public leaderboards (Chan et al., 2025), RE-Bench on research-engineering environments with 61 human experts as the reference (Wijk et al., 2024), and PaperBench on replicating published AI research under fine-grained rubrics (Starace et al., 2025). An outcome is the least arguable signal and the most expensive, and it folds method choice together with literature use, coding, and compute. MBM stops before execution and isolates two decisions inside method design, choosing an operating mechanism for a problem and then realizing it under that problem's constraints.

Scoring research by how good it looks. The opposite convention grades the proposal. In blind expert review, LLM-generated ideas were rated as more novel than expert ideas but less feasible (Si et al., 2025). Systems in this line retrieve and recombine literature to propose directions: SciMON scores generated ideas against prior papers and iterates until they are novel enough (Wang et al., 2024), ResearchAgent revises problem, method, and experiment design under feedback from LLM reviewing agents (Baek et al., 2025), and Co-Scientist generates, critiques, and evolves hypotheses for a scientist's stated objective (Gottweis et al., 2026). A rating collected before execution need not predict what execution reveals, and the Ideation–Execution Gap shows it directly: the same ideas, once carried out by experts, lose the advantage they held as proposals (Si et al., 2026). MBM's target is neither an outcome nor a rating, but agreement with a route a research group committed to under real constraints and then documented in enough detail to audit.

Learning the judgement, not only measuring it. Scientific Taste trains a judge on field- and time-matched pairs of high- and low-citation papers, then uses that judge as a reward for proposing ideas, which turns research judgement into a training target (Tong et al., 2026). Section 7 enters

the chain one step earlier and trains the ability to work out what a mechanism commits to, asking whether the choice that precedes that work improves with it.

Representing a method so that routes stay comparable. Architecture search automates method choice without representing it: the space is fixed in advance, so the decision becomes a search and what the candidates mean never has to be stated (Elsken et al., 2019). Routes across papers are not enumerable that way, so comparing them needs a unit that is neither a topic label nor a finished method, and several lines have converged on structured intermediate units. Scideator decomposes papers into purpose, mechanism, and evaluation facets and recombines them across papers (Radensky et al., 2024). MOOSE-Chem makes an inspiration a retrievable unit that composes with a research background into a hypothesis, and measures rediscovery against 51 papers published after the model's cutoff (Yang et al., 2025); ResearchBench divides the same task into retrieving inspirations, composing, and ranking, with candidate inspirations given as paper titles and abstracts (Liu et al., 2026). Matter to Mechanism annotates failure mode, intervention, mechanism, and target property as separate fields of a solution hypothesis in materials and battery research (Sourav et al., 2026). MBM adds one requirement to the unit: the same card must serve as a candidate for papers from different areas and stay specific enough to grade a realization against the source. That is what lets Allocation and Realization be asked about one paper, so route precision and core recall can separate a failure of choice from a failure of specification.

Where the negatives come from. What most separates these lines is the source of the negative example. Executed benchmarks take it from outcomes and pay for it by running the experiment. Open-ended ideation either has no negative or generates one, and a generated negative can leave a signature that makes the label recoverable from the option alone, which is how hypothesis-only models solved much of natural-language inference (Gururangan et al., 2018). Preference learning over matched publication pairs avoids generation by using two real papers, at the cost that the pair differs in many respects at once (Tong et al., 2026). MBM draws every alternative from a reviewed library of mechanisms that other papers did pursue and screens them for plausibility under the present problem's constraints, so an alternative is credible here without being correct here. Two further risks are answered in the construction, not in a list of caveats: a four-way prompt elicits option-position preferences independent of content (Zheng et al., 2024), so correct-option positions are balanced exactly, and a corpus drawn from published papers can sit inside pretraining (Oren et al., 2023), so answers are produced with no retrieval or web access and the card vocabulary is written to leave the mapping onto four abstract principles still to be done.

4 Constructing the paired benchmark

Construction converts a published method into a problem, an abstract mechanism, and a grounded realization reference. The goal is to preserve the method's design logic while controlling what the model sees. Figure 2 illustrates the relationship using a public-source reconstruction of LanczosNet (Liao et al., 2019); the full authoring and review flow appears in the appendix.

Both views come from one chain, not two questions attached afterwards. The three artifacts constrain one another: a card that cannot explain the realization is too loose, and a problem that reveals the binding gives the choice away.

Problem statements. A problem statement keeps what must be solved and under what constraints. It excludes the method, the observed outcome, and anything that identifies the paper. What remains is a decision that could be made before the experiment ran.

Mechanism cards. Each card names the object a mechanism acts on, the transformation it applies, the signal that guides it, and the conditions where it applies. Documented routes include optimal transport, Monte Carlo tree search, model-predictive control, ADMM consensus splitting, and conformal prediction sets—operating principles at a grain that recurs across papers and domains. The reusable library contains 243 reviewed cards, and Table 1 lists the routes that recur most often. Cards recur across papers by design: the same principle should be recognizable even when its architecture or domain changes.

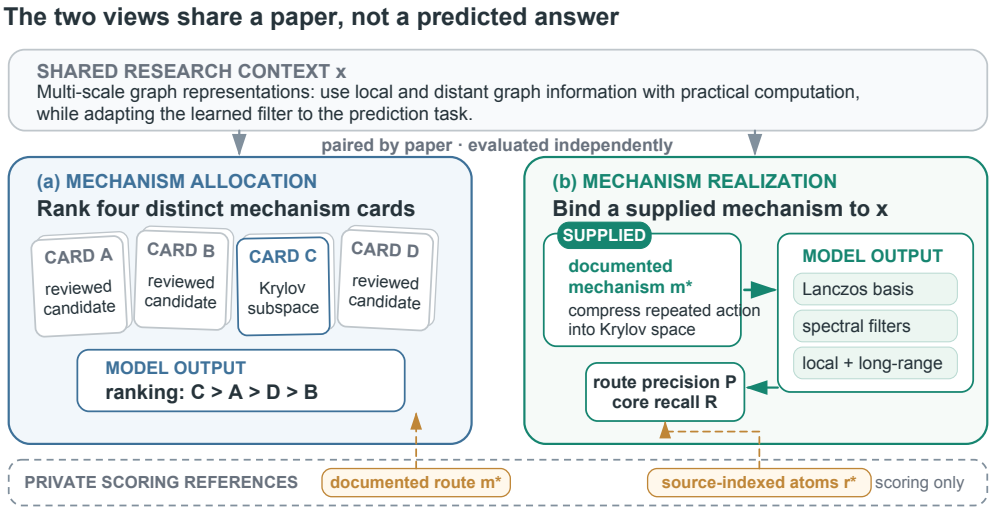

Figure 2: The two views of one paper. Both start from the same problem statement x: Allocation ranks four reviewed mechanism cards, while Realization receives the documented mechanism m^* and asks what it commits to in the target setting. No prediction passes between them. The documented route and the source-indexed reference atoms stay on the scoring side and are never shown to the model. The content paraphrases the public description of LanczosNet (Liao et al., 2019) and is not a benchmark item.

Choosing the alternatives. Each documented route is paired with three nearby mechanisms drawn from the reviewed card library. The alternatives are screened for plausibility under the problem's constraints and for mutual distinctness. No evaluated-model answer or score is used to choose an option.

Source-grounded realization references. References are drafted from the source paper by a model under a frozen protocol—the method must follow the card's principle, and the principle's instantiation for the problem must be identifiable in the source—and a random 10% of records is then confirmed by hand. The reference keeps only the design choices that must be specified before an experiment, omitting replaceable parameter settings and facts known only after execution. These choices are decomposed into 10,270 requirements, of which 4,415 are marked essential for the primary recall score. Matching accepts concise paraphrases and established near-equivalents, so atomization does not impose a fixed wording or component layout. A component is not a requirement atom: each answer is decomposed the way the reference is, so one component can satisfy several requirements.

Paired task views. Mechanism Allocation presents four abstract cards and requests a ranking. Mechanism Realization presents the same problem with the documented mechanism and asks for a small set of high-level components. A useful component says how the principle binds to this target rather than restating the principle. Table 2 in the appendix shows one complete Allocation item.

Composition and splits. MBM contains 789 papers, of which a fixed 300 form the evaluation split; every paper contributes both tasks to the same partition, and the remaining partitions are listed in the appendix. The corpus is recent: of the 735 papers with a known year, 343—nearly half—appeared in 2024 or later. The annotated evaluation set spans nine broad areas—learning methods at 31.7%, robotics and control at 16.7%, language and information at 14.7%, vision and media at 12.0%, and five smaller ones—across 35 finer directions, and the frequent documented routes enter that split in equal numbers (Appendix A).

Table 1: The fifteen most frequent documented routes over all 789 papers. These fifteen account for 67% of the corpus; the remaining third spreads over 87 further mechanism identities, so the label distribution has a long tail rather than a dominant class.

Mechanism	Papers	Mechanism	Papers
Evolutionary selection	71	Contextual bandit allocation	28
Optimal transport	64	Kalman and Bayesian filtering	26
Monte Carlo tree search	52	Thompson posterior sampling	24
Model-predictive control	42	EM latent alternation	21
ADMM consensus splitting	39	Swarm collective dynamics	20
Conformal prediction sets	37	Information bottleneck	19
Robust minimax	36	Spectral and Fourier methods	19
Control state space	33		

5 Scoring and evaluation protocol

Allocation uses Top-1 documented-route accuracy as its primary metric; Top-2 and reciprocal rank are secondary. Realization separates whether proposed content remains on the documented route from whether the answer covers its indispensable commitments. For paper i, let $z_{i\ell} \in \{0, \frac{1}{2}, 1\}$ be the alignment credit of proposed component ℓ, and $u_{ik} \in \{0, \frac{1}{2}, 1\}$ the coverage credit of core reference atom k. We compute

$$P_i = \frac{\sum_{\ell=1}^{L_i} z_{i\ell}}{L_i}, \qquad R_i = \frac{\sum_{k=1}^{K_i} u_{ik}}{K_i}, \qquad F_i = \frac{2P_iR_i}{P_i + R_i}, \tag{5}$$

Scores are computed per paper and then macro-averaged. Precision penalizes fluent but off-route additions; recall rewards essential coverage. Either alone is gameable: one safe statement keeps precision high, and scattering plausible details raises recall.

Realization answers are scored by GPT-5.6 Sol against source-indexed requirements, in two passes with targeted adjudication where the passes disagree. The judge receives one anonymous packet per paper with system identities and prior scores removed. It is itself one of the evaluated systems, and it places third on the metric it scores. Model judges carry position and self-enhancement biases (Zheng et al., 2023), which is what the anonymous packet and the order-reversed passes are there to contain. Scoring the ten answers together also makes a score depend on the set it was scored with, so these numbers describe this cohort and would have to be recomputed to admit a new system. The scores support comparisons between clearly separated systems rather than exact ranks; the audit is in the appendix.

Partial coverage carries half credit, and the setting does not drive the ordering: rescoring with no partial credit, or with three tenths, leaves the ten-system ranking unchanged and the same system first.

Position, card-length, and TF–IDF baselines remain near chance at 24–27% Top-1 (Appendix C).

Systems answer from the prompt alone, with no retrieval, tools, or web access, so the only route to a remembered answer is pretraining. Recalling a paper is a step short of answering it here, because the candidates are written in a target-independent card vocabulary, so a system must still map what it recalls onto one of four abstract principles. Systems reading the same items separate by nearly twenty points, which recognizing a source title could not produce on its own. Three variants of one model family span 5.83 points of Allocation accuracy between them, and to the extent that variants of a family share a pretraining corpus, exposure cannot be what separates them.

Memorization alone is bounded twice over by the corpus: correct options are balanced exactly across the four positions in the evaluation split, so label priors are worth 25%, and 226 of those papers appeared in 2025 or later, after the training cutoff of several evaluated systems, with the yearly count still rising at the end of the range.

Aggregating the same per-paper outcomes by pairwise Elo reproduces the leaderboard except for one adjacent pair that the two aggregations order differently, which is what treating adjacent ranks as indistinguishable already implies.

Table 2: Ten external systems on all 789 papers. Allocation reports documented-route Top-1; Realization reports route precision (P), core recall (R), and per-paper F1. Scores are percentages.

	Allocation	Mechanism Realization		
System	Top-1	P	R	F1
GPT-6 Astra	**69.20**	**65.99**	65.87	63.03
Claude Opus 5	66.54	63.87	**71.19**	**64.68**
GPT-5.6 Sol	62.48	64.93	62.85	61.15
Kimi K3	62.23	61.76	63.56	59.96
GLM 5.3	60.84	60.46	65.03	59.86
GPT-5.6 Terra	60.08	63.94	60.98	59.60
Claude Sonnet 5[†]	59.95	59.59	63.22	58.31
GPT-5.6 Luna	56.65	62.57	60.12	58.41
HY4 Preview	56.02	62.28	63.87	60.27
DeepSeek V4 Pro	49.43	60.01	59.68	56.76

External systems were not trained on MBM, so this table uses the complete corpus. Realization agreement and its audit are reported in Appendix B. [†]One abstention is retained as zero in the 789-paper denominator.

6 What current systems can design

Leaderboard and headroom. No system leads on both views. Astra leads Allocation and route precision; Opus leads core recall and Realization F1. Research design does not reduce to a single quality score.

Under the strongest system the documented route reaches the top two for 85.04% of papers but first for 69.20%, and a gap of fifteen to twenty-one points holds for every one of the ten systems (Appendix B). Systems reliably discard the candidates that do not fit the problem. What they cannot do is settle the last comparison between two routes that both look defensible.

A macro-averaged F1 near 65 reads better than the task has been solved. Counting papers whose answer covers at least 80% of the core commitments, the best system reaches 29.66% and the weakest 15.59%: on roughly seven papers in ten, even the leading system leaves the essential commitments incompletely stated.

6.1 Realizations help when they discriminate among research routes

Supplied realizations. A matched diagnostic scores the same 300 papers twice: once with abstract cards, once with cards carrying a prewritten realization. Every card's realization is built the same way: drafted and cross-revised by the same two strong models, one stronger on coverage and one on precision, and polished into the strongest version of that card's own route. None of the four is taken from the source paper's own realization, so the cards differ in the route they describe and not in how their realization was produced or how hard it was made to work. Allocation rises by 4.70 percentage points on average, and 10 of 11 systems improve, which a two-sided sign test puts at $p = 0.012$ (Figure 3; Appendix C). The weakest system on abstract cards gains the most, and the one regression is small. Concrete commitments add information an abstract card cannot: they show how a candidate would interact with the objects and constraints of the target problem.

Self-generated realizations. Fluent elaboration alone does not reproduce that benefit. On a matched 300-paper Qwen3.5-35B-A3B diagnostic, asking the model to draft its own realization before choosing lowers Allocation by 5.17 points. The drafts are plainly not ignored: shuffling them across candidates costs a further 13.11 points and leaves accuracy at 31.1%, against 25% chance. Training the selector to work with its own notes does not recover the loss either.

The model is thus highly responsive to implementation content and still worse off for having produced it.

A longer, busier prompt cannot carry the result on its own: the shuffled arm holds note content, length, and count fixed and changes only which note sits beside which card. Without notes the

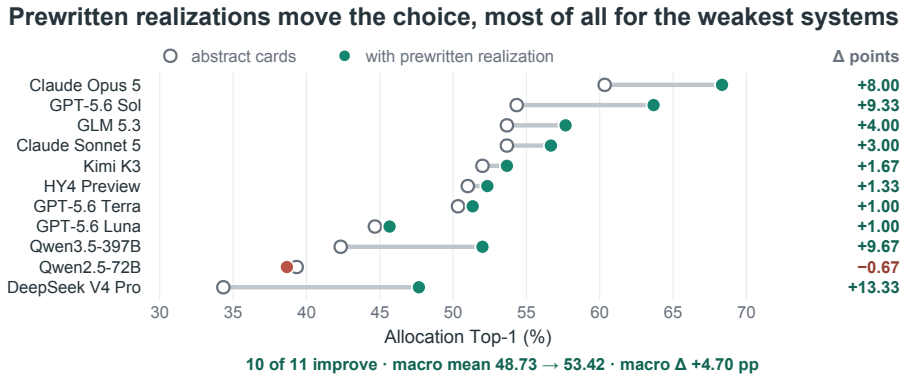

Figure 3: Each system answers one fixed 300-paper set twice, with abstract mechanism cards (open circles) and with cards carrying a prewritten realization (filled), ordered by abstract-card accuracy. The single regression is red. Qwen2.5-72B and Qwen3.5-397B belong to the common queue but not to Table 2.

selector scores 49.3%, and a mispaired note carries the choice most of the way down to the 25% floor.

That pattern fits a self-persuasion reading. Having drafted a workable implementation for every card, the model treats the coherence of its own draft as evidence about that card—and every reviewed card admits a coherent implementation, which leaves that evidence nearly constant across the four. An evaluator scoring its own text above an equally good alternative is documented (Panickssery et al., 2024), as is a loss of accuracy from deliberating where the deliberation is not what gets judged (Liu et al., 2024) and from revising one's own output without external feedback (Huang et al., 2024a). Here the drafted text is not the answer at all, only evidence for choosing one. Training the selector on its own notes narrows the loss without reliably closing it. We cannot separate this from miscalibration, where the notes carry binding information the selector weights badly: the two agree on every aggregate available here, and only the per-item choice transitions would tell them apart.

6.2 Allocation and realization are globally related but locally distinct

Across the ten systems in Table 2, Allocation and Realization rise together ($r = 0.857$ for F1). Among its two components the association is stronger for core recall (0.730) than for route precision (0.661): staying on route is a different capability from covering what the route requires. But a correlation across systems can be produced entirely by strong systems scoring well everywhere and easy papers being easy for everyone. In a panel where every system and every paper carries its own baseline, correct Allocation adds only 0.93 percentage points of Realization F1 on the same paper, with a 95% interval of $[0.25, 1.62]$ against a 7.92-point spread across systems. The cross-system correlations run from 0.661 to 0.857 over the four realization metrics while their same-paper coefficients all fall between $+0.66$ and $+0.94$, so the dissociation is a property of the two abilities and not of the metric used to measure realization. A system that realizes a mechanism faithfully on a given paper is barely more likely to have chosen that mechanism on the same paper: the two abilities travel together across systems and come apart within them. One number for research ability would average over exactly the gap this benchmark is built to expose, and training against one number would let a system buy its improvement on whichever side is cheaper.

7 Internalizing realization improves allocation

Realization-only reinforcement learning raises F1 from 37.60% to 40.98%, with route precision and core recall both improving (Appendix D). A short Realization phase before the same Allocation training then raises Allocation accuracy from 44.00% to 45.67%—five items on the 300 evaluation

papers, from a single run (Table 6). Realization is learnable, and some of what is learned carries into route choice.

At Allocation test time no realization is supplied: the model sees only the problem and candidate cards, so the gain cannot come from the earlier context intervention.

Two interventions on the same model and the same 300 papers then point in opposite directions. Supplying the model's own realization at the moment of choice costs 5.17 points. Training the model to produce realizations, and withholding every realization at test time, gains 1.67. What differs is not the information available when the choice is made—in the training arm there is none—but whether the ability to work out a mechanism's consequences sits in the weights. Realization appears to help allocation once it has been internalized, and to compete with the decision when it is injected as context.

A correlation between realization quality and route accuracy across ten systems admits a dull explanation—stronger systems are stronger at everything—and a substantive one: realization ability is part of what route choice draws on. The intervention favours the second, because moving realization by training moved allocation with it, on a model whose general strength was held fixed. The weak within-paper coefficient then says that whatever has been internalized is not yet deep enough to govern an individual decision.

Were Realization only a second scoreboard, a system could be tuned for it in isolation and the benchmark would measure two unrelated skills. A positive trace on later Allocation makes it a candidate intermediate objective instead: something to train on not because its own score matters, but because practising it moves a decision taken without it.

The system-level association is cross-sectional and the training result is one intervention; together they point the same way without establishing that a system better at realization is easier to train toward allocation. Testing that needs the same recipe across systems of differing realization strength, with several seeds and a budget-matched comparator.

8 Discussion and conclusion

Research design is not a single act of proposal generation but a chain of linked decisions: choose a scientific principle, work out its operational consequences, and use those consequences as evidence about whether the choice is sound. The mechanism representation makes this sequence measurable without invented bad methods or full experimental reproduction.

More implementation detail is not enough. The system that writes the best realizations is not the one that chooses best, and a system's own draft lowers the very choice it was meant to inform. What is missing is the set of obligations a principle carries in a specific setting, and their use in deciding. Both can be graded before an experiment runs, where a research decision is still cheap to change, and the learning target is therefore not a longer proposal but a realization whose commitments are reliable enough to carry the next decision.

Limits. Allocation measures agreement with a documented route, which is auditable but not a claim of unique optimality. The learning experiments use one model family and one completed run per recipe, so the transfer result marks a direction and not an effect size. Domain labels cover that split only. Scoring the realization side depends on a model judge whose repeat agreement is high for route alignment and lower for which requirements count as core, which is why we compare separated systems and not adjacent ranks.

AI USE STATEMENT

Generative AI tools executed parts of this research. They drafted the benchmark problems, mechanism cards, and realization references from the source papers under frozen protocols, screened the candidate alternatives, wrote the analysis and figure code, and served as the Realization judge that Appendix E describes separately from answer generation. The benchmark content is therefore model-authored and confirmed by hand on a sample rather than written from scratch. Tools also assisted with polishing the manuscript. The authors take responsibility for the final text, methodology, evidence, and artifacts.

ETHICS STATEMENT

MBM studies research-method decisions grounded in published work. Its labels describe documented routes rather than universal scientific merit; the scores should not be used to rank researchers or replace experimental validation. On acceptance we will release the benchmark in full and in public, including the source-indexed reference answers, with attribution to every source paper. Model-service terms and source licenses are respected in that release, and paper-derived material is redistributed in the abstracted form the benchmark uses rather than as reproduced source text.

REPRODUCIBILITY STATEMENT

The task definitions are specified in Section 2 and the scoring equations in Section 5. Appendix A details the paired partitions and construction stages; Appendix B specifies scoring denominators and statistical units, and Appendix D records the training endpoints. The manuscript tables and figures are generated from the recorded results by scripts, not transcribed by hand. Every prompt sent, answer returned, and answer scored is kept with the run that produced it, so no reported number can be revised after the fact. The supplementary material accompanying this submission contains ten Allocation items exactly as a system receives them, with the correct option unmarked, both system prompts, and the stated rule that selects those ten, which can be rerun against the full set once it is released. On acceptance we will publicly release all of it: the model-facing problems and mechanism cards for all 789 papers, the source-indexed realization references and scoring rubrics, the scoring and analysis code, and the recorded outputs of every system reported here. Reference answers are withheld from the evaluated systems at answer time, not from the community after publication; the numbers reported here therefore describe systems evaluated before any public release.

REFERENCES

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, 2025. URL https://arxiv.org/abs/2404.07738.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021. URL https://aclanthology.org/2021.naacl-main.385/.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07095.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*,

2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21, 2019. URL https://arxiv.org/abs/1808.05377.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2): 155–170, 1983. doi: 10.1207/s15516709cog0702_3.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL https://doi.org/10.1038/s41586-026-10644-y.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018. URL https://aclanthology.org/N18-2017/.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024a. URL https://arxiv.org/abs/2310.01798.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20271–20309. PMLR, 2024b. URL https://proceedings.mlr.press/v235/huang24y.html.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 375–385. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Renjie Liao, Zhizhen Zhao, Raquel Urtasun, and Richard S. Zemel. LanczosNet: Multi-scale deep graph convolutional networks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1901.01484.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse, 2024. URL https://arxiv.org/abs/2410.21333.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. ResearchBench: Benchmarking LLMs in scientific discovery via inspiration-based task decomposition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13187–13207, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.644. URL https://aclanthology.org/2026.findings-acl.644/.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651(8107): 914–919, 2026. doi: 10.1038/s41586-026-10265-5. URL https://doi.org/10.1038/s41586-026-10265-5.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models, 2024. URL https://arxiv.org/abs/2407.01725.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black box language models, 2023. URL https://arxiv.org/abs/2310.17623.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations, 2024. URL https://arxiv.org/abs/2404.13076.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. Human–LLM compound system for scientific ideation through facet recombination and novelty evaluation, 2024. URL https://arxiv.org/abs/2409.14634v7. Version 7 reviewed, revised 2026; the system is named Scideator.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ea94957d81b1c1caf87ef5319fa6b467-Abstract-Conference.html.

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The ideation–execution gap: Execution outcomes of LLM-generated versus human research ideas. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Fllp8l6Puy.

Shashwat Sourav, Tanjin. He, Maria K. Y. Chan, Anubhav Jain, and Tirthankar Ghosal. Matter to mechanism: A benchmark for AI co-scientists in materials and battery research, 2026. URL https://arxiv.org/abs/2606.02258.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's ability to replicate AI research, 2025. URL https://arxiv.org/abs/2504.01848.

Jingqi Tong, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Zhiheng Xi, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, Ming Zhang, Qiguang Chen, Weifeng Ge, Qipeng Guo, Tianlei Ying, Tianxiang Sun, Yining Zheng, Xinchi Chen, Jun Zhao, Ning Ding, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. AI can learn scientific taste, 2026. URL https://arxiv.org/abs/2603.14473v2. Version 2 reviewed for the positioning claims.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://aclanthology.org/2024.acl-long.18/.

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Holden Karnofsky, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes. RE-Bench: Evaluating frontier AI r&d capabilities of language model agents against human experts, 2024. URL https://arxiv.org/abs/2411.15114.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search, 2025. URL https://arxiv.org/abs/2504.08066.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. MOOSE-Chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07076.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.03882.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.05685.